Etymological studies are generally published in lexicographic prose, relying on bibliographies for cross-referencing (see, for examples, Mallory & Adams, 2006, and Beekes & van Beek, 2009). Authoritative sources tend to be available in unstructured formats only, with on-line resources in most cases either unstructured, commendable but limited in scope or intended for phylogenetics, criticized in their theoretical assumptions, or academically unsuitable. Extensive labor is necessary for aggregating information such as competing reconstructions, imposing limits on data sharing and comparison (Forkel et al., 2018; Rzymski, Tresoldi et al., 2020).

*CompIE* is a proposal for a database cross-referencing authoritative Indo-European comparative material. It orients towards inter-operable and open data by adopting tools from the CLDF Initiative (Forkel et al., 2018; Forkel et al., 2019) and the relational model, with data as tuples of relations defined in consistence with first-order predicate logic (Codd, 1970). Information is collected in textual long-table format, editable with common spreadsheet programs and suitable for documenting changes. Entries are cross-referenced with internal and external catalogs (Anderson et al., 2018; List et al., 2016; Hammarström et al., 2019; Rzymsky, Tresoldi et al., 2020), allowing composite queries, and data validation. The emphasis on referencing and the mandatory bibliographic field orients the proposal as a tool for linking, and not replacing, authoritative data (Wilkinson et al., 2016).

The core of the database are collections of cognates linked through identifiers, listing sources where scholars can obtain additional, supporting, or opposing material:

<table>
<thead>
<tr>
<th>ID</th>
<th>PIE-ID</th>
<th>Language</th>
<th>Form</th>
<th>Segments</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT32</td>
<td>PIEyoke</td>
<td>Hittite</td>
<td>𒄿𒌑𒃷</td>
<td>juka</td>
<td>DeVaan08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Attic</td>
<td>ζυγόν</td>
<td>zdygón</td>
<td>DeVaan08,Liddell40</td>
</tr>
<tr>
<td>AT567</td>
<td>PIEyoke</td>
<td>Proto_German</td>
<td>*jukâ</td>
<td>jukâ</td>
<td>Kroonen13</td>
</tr>
<tr>
<td>PG9</td>
<td>PIEyoke</td>
<td>Imperial_Latin</td>
<td>jugum</td>
<td>jugum</td>
<td>DeVaan08</td>
</tr>
<tr>
<td></td>
<td>PIEyoke</td>
<td>Portuguese</td>
<td>jugo</td>
<td>zuyu</td>
<td>Aurelio87</td>
</tr>
</tbody>
</table>

By adopting meaningless unique identifiers, the proposal takes an agnostic view in face of competing information and facilitates internal catalog referencing:
Internal reconstruction is accounted for by extending models for the annotation of word-formations, still with partial support for processes such as apophonies and reduplication:

The proposal extends sequence alignment, as used in phylogenetics for automatic cognate detection (List, 2014, Kilani, forth), to support non-phonological alignment, with algorithms for identifying patterns of correspondence that experts can correct. The segmented sequences can later be filtered with multi-tiered representations:

Multi-sequence alignments of any kind of corresponding sequence are possible, including orthographies, reconstructions, and pronunciations:

along with partial cognacy alignment:
The long-table format facilitates the referenced and structured record of alternative information, such as competing etymologies, once more cross-linking etyma and glosses:

<table>
<thead>
<tr>
<th>PIE Concept</th>
<th>Etymon</th>
<th>Gloss</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>*hₐṛt-ko-</td>
<td>BEAR</td>
<td>*h₂rétk-</td>
<td>destroyer MalloryAdams: 55</td>
</tr>
<tr>
<td>*hₐṛt-ko-</td>
<td>BEAR</td>
<td>*h₂rtkō-</td>
<td>(sharp-)pawed GJP</td>
</tr>
</tbody>
</table>

A proof-of-concept of the proposal has been developed internally, with automatic deployment to a website when changes are committed, and is being used for data exploration, notably for morphology. At the time of submission it counts circa 2,500 Proto-Indo-European entries, 10,000 reflexes in over 100 languages, along with dozens of morphological annotations and alignments. It will be made available before presentation as a public website offering downloadable data.

References


